DOCUMENT RESUME

ED 317 867 CE 054 750

AUTHOR Sticht, Thomas G.
TITLE Testing and Assessment in Adult Basic Education and English as a Second Language Programs.
INSTITUTION Applied Behavioral & Cognitive Sciences, Inc., San Diego, CA.
SPONS AGENCY John D. and Catherine T. MacArthur Foundation, Chicago, IL.; Office of Vocational and Adult Education (ED), Washington, DC. Div. of Adult Education and Literacy.
PUB DATE Jan 90
NOTE 50p.
PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Accountability; *Adult Basic Education; Adult Education; Competency Based Education; Criterion Referenced Tests; *Educational Diagnosis; *Educational Legislation; Educational Policy; Educational Testing; *English (Second Language); Government School Relationship; Item Response Theory; Norm Referenced Tests; *Politics of Education; Predictive Validity; *Standardized Tests; Test Construction; Test Use
IDENTIFIERS Adult Basic Learning Examination; *Adult Education Amendments 1988; Basic English Skills Test; Comprehensive Adult Student Assessment System; English as a Second Language Oral Assessment; General Educational Development Tests; Reading Evaluation Adult Diagnosis; Tests of Adult Basic Education

ABSTRACT

This document expands upon the discussion of standardized tests in federal legislation and Department of Education rules and regulations, in order to guide practitioners in using these tests and alternative assessment methods more wisely. Amendments of 1988 that address the uses of standardized tests, the federal regulations that implement the amendments, public comments on the regulations, and the U.S. Department of Education's responses to the comments are presented. Chapter 2 deals with the nature and uses of standardized tests, including definitions of standardized tests, norm-referencing, criterion-referencing, competency-based education, and curriculum-based tests. Chapter 3 provides information about eight standardized tests in wide use in adult basic education (ABE) and English-as-a-Second-Language (ESL) programs: Adult Basic Learning Examination (ABLE); Basic English Skills Test (BEST); Comprehensive Adult Student Assessment System (CASAS) ABE and ESL tests; English-as-a-Second-Language Oral Assessment (ESLOA); General Educational Development Official Practice Tests; Reading Evaluation Adult Diagnosis (READ); and Tests of Adult Basic Education (TABE). Chapter 4 discusses "negative gain" scores, general and specific literacy, item response theory, predictive validity, special problems in testing in ESL programs, alternative assessment methods, and assessment systems to meet instructional purposes and state and federal requirements for accountability. Thirty-three reference footnotes are given. Appendix A provides a table for comparing scores among several standardized tests. Appendix B provides 13 sources of additional information and a set of 5 transparency masters for use in presentations on standardized testing in ABE and ESL programs. (CML)

ED317867

Testing and Assessment

in
Adult Basic Education
and
English as a Second Language Programs

Thomas G. Sticht
Applied Behavioral & Cognitive Sciences, Inc.
San Diego, CA

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

January 1990

Preparation of this report was supported by the U. S. Department of Education, Division of Adult Education and Literacy, and by a grant from the John D. and Catherine T. MacArthur Foundation to the Applied Behavioral & Cognitive Sciences, Inc.

CE 059 750

BEST COPY AVAILABLE

FOREWORD

This report on testing and assessment was prepared for the Division of Adult Education and Literacy, Office of Vocational and Adult Education by Dr. Tom Sticht. Our purposes for distributing it are to encourage discussion of assessment issues and to promote understanding of tests and their use. The report should also be useful as a reference and as a resource for both ABE and ESL testing. Because this document reflects the author's professional judgement, it does not represent positions or policies of the U.S. Department of Education, and no official endorsement should be inferred.

Acknowledgments

Preparation of this report was supported by the U. S. Department of Education, Division of Adult Education and Literacy, and by a grant from the John D. and Catherine T. MacArthur Foundation to the Applied Behavioral & Cognitive Sciences, Inc.

James Parker was project officer at the U. S. Department of Education and provided much guidance that helped shape the contents and format of the report. His many contributions are greatly appreciated.

Several colleagues reviewed early drafts of the report and provided comments helpful for its improvement. Thanks for this valuable assistance are due and hereby given to Bill Armstrong, Bill Grimes, Nancy Hampson, Timothy Houchen, Leann Howard, Michael O'Malley, Ron Pugsley, Patricia Rickard, Dick Stiles, and Ruth Weinstock.

The Association for Community Based Education (ACBE) generously permitted Dr. Gregg Jackson to prepare the insightful test reviews of Chapter 3 by drawing upon materials from a much more extensive review of tests that he wrote for the ACBE. Grateful appreciation is expressed to both the ACBE and to Dr. Jackson for their important contributions to this report.

The assistance from the foregoing, and others who contributed pre-prints, unpublished data, and other resources is gratefully acknowledged. However, it should be noted that any remaining limitations, shortcomings, or errors of omission or commission are due to the author.

Similarly, the findings and opinions expressed in this report are solely those of the author, and they do not necessarily represent the opinions, policies, or official positions of the U. S. Department of Education, the John D. and Catherine T. MacArthur Foundation, or the Applied Behavioral & Cognitive Sciences, Inc.

About the Author

Dr. Thomas G. Sticht is President and Senior Scientist, Applied Behavioral & Cognitive Sciences, Inc., San Diego, CA. Previous positions include Vice President and Director of the Basic Skills Division, Human Resources Research Organization (HumRRO), Alexandria, VA; Associate Director of the National Institute of Education, a Senior Executive Service level position in the former U.S. Department of Health, Education, and Welfare; and a faculty member of Harvard University, American University, the U.S. Naval Postgraduate School, and community colleges in California and Virginia. He is a member of the Professional Advisory Board of the Business Council for Effective Literacy, the National Commission on testing and public policy, and chair of the California Workforce Literacy Task Force (1989-91).

Dr. Sticht directed HumRRO's functional literacy R&D program that produced the Army's landmark Functional Literacy (FLIT) program. In this and subsequent work, he directed the development of job-related literacy programs, including the development of job-related reading tests for the U.S. Air Force, Army, and Navy using traditional and item response theory-based psychometrics. He is the author of over a hundred reports, articles, and books on adult literacy, workplace literacy, and intergenerational literacy. His Ph.D. is in experimental psychology from the University of Arizona.

Table of Contents

Chapter 1	Standardized Testing in ABE and ESL Programs	1
Chapter 2	Nature and Use of Standardized Tests	5
Chapter 3	Review of Tests for ABE and ESL Programs	13
Chapter 4	Special Topics in the Use of Standardized Tests	22
Footnotes		30
Appendix A	Correspondences Among Frequently Used Tests	33
Appendix B	Resource Materials	37

Chapter 1

Standardized Testing in ABE and ESL Programs

The Adult Education Act, Amendments of 1988, require that States evaluate at least one-third of the recipients of grants for adult basic education (ABE) and English as a Second Language (ESL) programs. These evaluations are to include, among other things, data from the use of standardized tests.

This report expands upon the discussion of standardized tests given in the federal law and the Department of Education rules and regulations that implement the law. It provides information that can be helpful to practitioners in selecting and using standardized tests. It may also serve as a resource for staff development. It is hoped that the discussion of concepts, issues, and definitions will help program administrators and teachers to more wisely use standardized tests and alternative assessment methods for program evaluation. To this end, topics such as *reliability* and *validity* are discussed in the context of specific problems providers frequently face, rather than as separate psychometric concepts.

OVERVIEW OF THE REPORT

Following the Introduction to and Overview of this report, Chapter 1 presents a summary of the Amendments that address the uses of standardized tests, the federal regulations that implement the Amendments, the public comments on the Department of Education regulations, and the Department's response to those comments. This summary reveals some of the issues surrounding the uses of standardized testing in adult education. Additionally, it calls attention to technical terminology and other aspects of standardized testing that may be unfamiliar to many who are presently or about to be involved in ABE or ESL program development and implementation.

Chapter 2 deals with the nature and uses of standardized tests. The purpose is to elaborate on the federal definition and discussion, so that users of standardized tests in adult education programs will have a better understanding of what standardized tests are and how to use them appropriately. This section answers questions such as, What does it mean to say that a test is standardized? What is a norm-referenced test? What is a criterion-referenced test? What is competency-based education and how does it relate to the use of norm- or criterion-referenced tests? What is a curriculum-based test?

Chapter 3 provides information about eight standardized tests (or test systems) that are in wide use in ABE and ESL programs. These include the Adult Basic Learning Examination (ABLE), the Basic English Skills Test (BEST), the Comprehensive Adult Student Assessment System (CASAS) ABE and ESL tests, the English as a Second Language Oral Assessment (ESLOA), the GED Official Practice Tests, the Reading Evaluation Adult Diagnosis (READ) test, and the Tests of Adult Basic Education (TABE).

Chapter 4 discusses special topics in the use of standardized tests, including: What to do about "negative gain" scores, that is, when students do poorer at the end of the program than they did at the beginning? What is the difference between "general" and "specific" literacy and when should programs assess each? What is predictive validity and what does it have to do with assessment in ABE and ESL programs? How does a test that is developed using item response theory differ from traditional tests? What are some special problems in testing in ESL programs? What are "alternative assessment" methods? What kind of assessment system can be developed to meet instructional purposes and State and federal requirements for accountability?

Appendix A presents a table for comparing scores among several standardized tests, while Appendix B provides sources of further information, and a set of transparency masters that may be used in presentations on standardized testing in ABE and ESL programs.

STANDARDIZED TESTING IN ADULT EDUCATION

The Adult Education Act, as amended in 1988, requires State adult education agencies to "gather and analyze data (including standardized test data) to determine the extent to which the adult programs are achieving the goals set forth in the [State] plan..."[1]

In implementing the Adult Education Act, the U. S. Department of Education Rules and Regulations for evaluating federally supported State Adult Education Programs requires that State Education Agencies "gather and analyze data on the effectiveness of all State-administered adult programs, services, and activities - including standardized test data...."[2]

Public Comments on Standardized Testing in ABE and ESL Programs

Public comments on the requirements for standardized test data in the federal law and regulations for adult education raised a variety of issues regarding standardized tests and their uses in assessing students and programs.[2] The comments included:

> The provision concerning standardized test data should be deleted or made optional.
>
> Should every student provide standardized test data?
>
> The regulations preclude better or more effective assessment methods, and other criteria should be used to determine program effectiveness.
>
> Measures of accountability are needed but they should be brief and non-threatening to learners.
>
> Students should be assessed in terms of skill strengths and weaknesses rather than by grade level norms.

> Questions were raised about benchmarks to determine program effectiveness, criteria or guidelines to be used in the selection of assessment instruments for special populations in adult education, and standards for personnel who administer assessment in the field.
>
> The U.S. Department of Education should identify a variety of standardized tests or specify which assessment instruments are appropriate for the various populations of adults.

Federal Statement on Standardized Tests

In response to the public comments summarized above, the U. S. Department of Education offered a definition of a "standardized test" and replies addressing the comments.

> Definition. A test is standardized if it is based on a systematic sampling of behavior, has data on reliability and validity, is administered and scored according to specific instructions, and is widely used. A standardized test may be norm-referenced or criterion-based. The tests may, but need not, relate to readability levels, grade level equivalencies, or competency-based measurements.[2]

Selection and Use of Standardized Tests. The federal response to the comments on standardized tests goes on to state that:

> It is inappropriate for the Department of Education to select or approve standardized tests, and that State Education Agencies must select the tests they deem appropriate for their State.
>
> A State need not have standardized test data for every student.
>
> States are required to gather and analyze data in addition to standardized test data to determine program effectiveness.
>
> States have flexibility to determine which criteria measure effectiveness within the federal framework for reviews and evaluations.

As the foregoing illustrates, adult education programs are required by federal law to provide standardized test data as one indicator of program effectiveness. But there are serious concerns from programs, with some suggesting that all standardized testing be dropped or made optional, and others suggesting that the federal government should indicate which standardized tests to use to evaluate programs.

For many adult educators, concepts such as "standardized," "norm-referenced," criterion-referenced," and others that appear in the federal regulations may be little understood. These and other concepts related to testing are discussed in Chapter 2 to provide adult educators with a better basis for making choices in response to State and federal evaluation requirements.

Chapter 2

Nature and Use of Standardized Tests

As the federal regulations note, a *standardized* test is a test that is administered under standard conditions to obtain a sample of learner behavior that can be used to make inferences about the learner's ability. A standardized test differs from an informal test in that the latter does not follow a fixed set of conditions. For instance, in a standardized reading test, the same reading materials are read by different learners following the same procedures, answering the same types of questions and observing the same time limits.

The purpose of the standard conditions is to try to hold constant all factors other than the ability under study so that the inference drawn about that ability is *valid*, that is, true or correct.

Standardized tests are particularly useful for making comparisons. They let us compare a person's ability at one time to that person's ability at a second time, as in pre-and post-testing. They also permit comparisons among programs. However, for the tests to give valid results for making such comparisons, they must be administered according to the standard conditions.

By understanding the logic of *standardization* in testing, programs can strive to keep the conditions of test administration from affecting test performance. Here are some things to avoid:

Avoid: Ignoring time standards. Here is a simple illustration of the reasoning behind the methodology of standard conditions. If a program wanted to compare a group of learners' post-program reading ability to their pre-program ability, and it only gave them fifteen minutes to complete a hundred items on the pre-test, then it would not be appropriate to let them have thirty minutes to complete a comparable set of items at the post-test. Using such different conditions of test administration, one could not infer that the learners' greater post-test scores indicated a true gain in ability over the pre-test scores. It might simply indicate that the learners were able to complete more items because there was more time. In this case, then, the learners' abilities had not increased. Rather, the conditions under which the test was administered were changed. They were not standard for both the pre- and the post-tests. And these changed conditions of administration may have produced the observed increase in test scores.

Avoid: Testing the first time students show up for a program. Many adult students will not be very comfortable at the first meeting. They may be nervous and frightened about taking a test. They may also be unprepared in test-taking strategies. Because of this psychological condition of the learner, they do not meet the conditions of standardization of most tests, which assume a more-or-less relaxed, test-experienced learner. If pre-tested under their first meeting psychological conditions, learners' true abilities may be greatly underestimated. Then, at the post-test, after they have had time to adjust to the program, its staff, and have had practice in answering test questions similar to the standardized tests, their post-test scores may be higher. But in this case, much of

the gain may represent the change in the learners' emotional conditions, and not gain in the cognitive ability (e.g., reading, writing, mathematics) that is the object of assessment.

The increase in post-test scores over pre-test scores due to the kinds of psychological factors discussed are sometimes called "warm-up," "surge" or "practice" effects. Such effects may be particularly troublesome when pre- and post-testing are separated by only a few hours. Some programs may have capitalized on such effects in claiming to make one, two or more "years" gain in reading or mathematics in just 15 or 20 hours of instruction.

In general, pre-testing should not be accomplished until learners have had an opportunity to adjust to the program and practice their test-taking skills.

TYPES OF STANDARDIZED TESTS

Scores on standardized tests do not have much meaning in and of themselves. If a learner correctly answers 60 percent of items on some standardized test, it is not clear what that means in the absence of other information that helps us *interpret* the score. We do not know if 60 percent indicates high ability or low ability in the domain being assessed (for example, reading). For instance, if every other adult similar to the learner scores 90 percent correct, then we would probably conclude that 60 percent was an indicator of low ability. To interpret the score, we need other information to which the observed score can be *referenced* or *based,* that is, compared and related.

The federal rules and regulations note that standardized tests may be norm-referenced, criterion-based, or competency-based. But it is not always clear just what different scholars or practitioners mean by these terms. The following discussion is meant to provide a common frame of reference for program operators for understanding the various types of standardized tests that are available.

Norm-Referenced Tests

All human cognitive ability is socially derived. That is, the language one uses, the concepts used for thinking and communicating, the logic of reasoning, the types of symbols and symbolic tools (e.g., tables, graphs, figures, bus schedules, tax forms, etc.), and the bodies of knowledge stored in people's brains or in books are developed by individuals being reared in social groups.

Because of the social basis of cognition, many standardized tests have been developed to permit a learner's score to be interpreted in relation to, or, stated otherwise, in *reference* to the scores of other people who have taken the test. In this case, then, an individual's standardized test score is interpreted by comparing it to how well the referenced group *normally* performs on the test. If the individual learner scores above the average or norm of the referencing or *norming* group, the person is said to be above average in the ability of interest. If the learner scores below the average of the referencing group, he or she is said to be below average in the ability.

Grade level norms. In adult literacy education programs, standardized tests are frequently used that have been normed on children in the elementary, middle, and secondary school grades. In this case then, the adult learner's score on the test may be interpreted in reference to the average performance of children at each grade level. If an adult's score on a reading test normed on grade school children is the same as that of a child in the eighth month of the fourth grade, the adult would be assigned an ability level of 4.8. If the adult's score was the same as the average for school children in the sixth month of the ninth grade, the adult would be said to be reading at the 9.6 grade level.

Interpreting these grade level scores for adult learners is not straightforward. For instance, the score of 4.8 does not mean literally that the adult reads like the average child in the eighth month of the fourth grade. In fact, in one research study adults reading at the fifth grade level were not as competent at other reading tasks as typical fifth grade children.[3] This is not too surprising when it is considered that the child is reading at a level that defines what is typical for the fourth grader, while the adult in our relatively well-educated and literate society who reads at the fourth grade level is well below the average for adults.

What the fourth grade score for the adult means is that the adult reads very poorly relative to other adults who may score at the ninth, tenth, or twelfth grade levels on the test. While the grade level score is based on the performance of children in the school grades, the interpretation of the score should be based on the performance of adults on the test. For this reason, standardized tests such as the Tests of Adult Basic Education (TABE) or Adult Basic Learning Examination (ABLE) provide norms for adults in adult basic education programs and other settings that permit test users to interpret scores both in grade levels (grade-school referenced norms) and in relation to adult performance on the tests.

Identifying differences among readers. The major use of norm-referenced test scores is to identify differences among a group of people for some purpose. The norm-referenced tests indicate how people perform relative to the norming group. For instance, are they below or above the average of the norming group.

The most widely used standardized, basic skills (reading, mathematics) test that is normed on a nationally representative sample of young adults (18 to 23 years of age) is the Armed Forces Qualification Test (AFQT). This test has been specially designed to permit the armed forces to rank order young adults from those very low to those very high in basic skills.and to screen out the least skilled from military service. The U. S. Congress has passed a law prohibiting young adults who score below the tenth percentile on the AFQT from entering military service.

Adult education programs frequently use norm-referenced reading tests to identify those with reading scores below the fourth or fifth grade levels, those scoring between the fifth and ninth grade levels, and those scoring at or above the ninth grade level. These categories are frequently used to assign adults to different levels of reading instruction: basic or beginning reading, mid-level reading, and high school equivalency (General Educational Development - GED) education.

The use of standardized, norm-referenced tests for selection or placement is not an altogether accurate procedure, if for no other reason than the fact that no test is perfectly *reliable*. That is, because of the differences in people's psychological conditions from time to time, and variations in the physical conditions of testing (for example, it may be very cold, or too hot, or too noisy one day, and so forth), people do not usually score the same on tests from one time to the next.

Also, when multiple-choice tests are used that have been designed to discriminate among a wide-range of ability levels, the tests will contain some very easy items, some average difficulty items, and some very difficult items. The multiple-choice format permits guessing. These conditions mean that a person may score correctly on some items by chance alone on one day, but not the next. This produces artifacts that should be avoided in adult education program evaluation.

Avoid: Regression to the mean. Because of the imperfect reliability of tests as discussed above, a phenomenon that has plagued adult education programs for decades is regression to the mean. This usually happens when a group of adults is administered as a pre-test, a standardized test that has been normed using traditional test development methods, and a part of the group is identified as low in ability and sent to a program. Then, later on, when just the low group is post-tested, it is found that the average post-test score is higher than the pre-test score. Under these circumstances, the program offers the gain between pre and post-test scores as evidence of the effectiveness of the program in bringing about achievement.

However, regression to the mean is a statistical process that generally operates under the foregoing conditions. Whenever a low-scoring group is separated off from the total group and then retested, the average score of the post-test will generally be larger than the average score of the pre-test. This is due to the fact that many people are in the low group on the pre-test because they guessed poorly or did not perform well due to anxiety, lack of recent practice in test-taking and so forth, as mentioned earlier. So, when they are retested, their average score moves up toward (that is, regresses toward) the mean (or average) score of the total group on which the test was normed. [4]

Such warm-up and regression effects can be quite large. In one study, military recruits new to the service were tested with a standardized, grade-school normed reading test. Those scoring below the sixth grade level were retested two weeks later, with no intervening reading instruction, and those who scored above the sixth grade were excluded from the study. Two weeks later, the remaining recruits who scored below the sixth grade level were retested with a third form of the reading test, and those who scored above the sixth grade level were excluded. This process reduced the number of people reading below the sixth grade level by 40 percent! [5]

Regression effects can be reduced in several ways. One is to use the retesting procedure discussed above. Obviously, this requires quite a commitment to testing. It also requires the use of standardized tests with at least three comparable forms, one for the first testing, a second for the next testing of the group identified as low on the first testing, and a third for the post-testing of the group identified in the second testing who were placed in the program of interest.

Regression effects can also be reduced by not testing learners until they have adjusted to the program and obtained some practice in test-taking as noted earlier.

In another approach to managing regression effects, scores on post-tests may be adjusted for regression by using the correlation between pre and post-test scores. This permits the prediction of post-test scores from pre-test scores. Then, actual post-test scores can be compared to the predicted scores. Only the gain that exceeds the predicted post-test scores is then used to indicate program effectiveness. This procedure requires technical assistance from a knowledgeable statistician or psychometrician.

Regression effects may also be estimated and adjusted for by comparing the program group to a group with similar pre-test scores which does not receive the educational program being evaluated (though note that the control group should receive some practice in test-taking, to offset the "warm-up," "surge" or "practice" effects discussed above). This "treatment" and "no treatment" groups comparison permits programs to adjust their gains for regression.

Use of tests with very low probabilities for guessing can also reduce regression. This will be discussed later on in regard to the problem of "negative gain."

Criterion-Referenced Tests

The concept of criterion-referenced assessment was stated in contemporary form by Glaser and Klaus.[6] The concept was advanced as a contrast to the wide-spread method of grading in educational programs known as grading "on the curve." In grading based "on the curve," learners' grades depend on how well everyone in the class or other norming group performs. An individual learner's grade is determined in relation to the grades of others. Therefore, if everyone in the class performs poorly, a low mark, say 60 percent correct, may be assigned a relatively high grade, say, a "B." Yet, if everyone performed well, a mark of 60 percent correct might be assigned a grade of "D."

In criterion-referenced testing, an absolute standard or criterion of performance is set, and everyone's score is established in relation to that standard. Thus, 90 percent correct and above might be necessary to receive a grade of "A," 80 to 89 percent correct for a "B," and so forth. In criterion-referenced testing then, learners' achievement in an instructional program is assessed in terms of how well they achieve some absolute standard, or criterion of learning, rather than by comparison to a norming group.

Using a norm-referenced test is like grading "on the curve." If the norming group improves overall, then tests may be renormed to adjust the average score higher. There will always be somebody below average. This does not permit one to say, then, how well someone has or has not mastered some body, or as it is called in test development, some domain of knowledge or skill.

Criterion-referenced testing had its roots in the behavioral psychology of the 1950's and 1960's, and was closely related to the development of self-paced, individualized, more-or-less carefully pre-programmed instruction. In instructional programs following this approach, a domain of knowledge and skill is carefully defined. Learning objectives that can be assessed are specified, and units of instruction, frequently called "modules" are developed to teach the various subsets of knowledge and skill identified by the learning objectives.

With the modules in place, learners are introduced to a module preceeded by a pre-module test, to see if they already know the material to some pre-determined criterion, e.g., 90 percent correct. If the learners pass the pre-module test, they go on to the next module with its pre-module test and so forth. If a pre-module test is failed, then the learner is assigned the study materials and lessons of the module in question, and then is administered a post-module test to see if he or she can perform at the desired criterion.

In this criterion-referenced approach to assessment, learner gain is interpreted in terms of how many units of instruction are mastered at the prescribed criterion level and not in terms of the learner's change relative to a norming group.

Competency-Based Education and Testing

Closely related to the concept of criterion-referenced testing is the concept of "competency-based" education. Just as criterion-referenced testing was put forth in opposition to the practice of grading "on the curve," a practice which obscures just how much learning may take place in a program, the concept of competency-based education was put forth in opposition to the traditional practice of awarding educational credit or certification on the basis of hours of instruction or number of courses completed. Such factors do not reveal the actual competence developed in the program of instruction.

The major factor distinguishing "competency-based" from "traditional" education is the idea that a learner's progress in the course should be based on the demonstration that new competence has been achieved, not on the basis of the number of hours or courses in which the learner has participated.

Because competency-based programs typically identify learning objectives very specifically, they tend to use criterion-referenced assessment. Sometimes, both criterion- and norm-referenced tests are used in competency-based programs. For instance, in the Job Corps program, or its "civilian" adaptation, the Comprehensive Competencies Program (CCP), a norm-referenced test, such as the TABE, is administered as a pre-test to determine the learner's general level of skill for placement into the instructional modules of the program. Then criterion-referenced assessment is used to indicate whether or not learners are mastering the specific course competencies, as in the pre- and post-module assessments mentioned above. Finally, norm-referenced, post-course tests are used to indicate growth in the "general" ability to which the specific competencies contribute.[7]

What makes the course "competency-based" is the fact that criterion levels of achievement on the norm-referenced tests are established, such as achievement of the 8th grade level, before promotion is made to the next level of education, such as high school equivalency instruction. The 8th grade level of achievement is the criterion that must be achieved for promotion to the next level of instruction. As this illustrates, norm-referenced tests may be used as criterion-referenced tests in competency-based instruction.

In the Comprehensive Adult Student Assessment System (CASAS) hundreds of basic skills (listening; reading; mathematics) competencies judged to be important to be mastered by adult basic education learners have been identified. For each of the hundreds

of competencies, a number of test items have been developed to assess mastery of the competencies at different levels of difficulty. These thousands of test items have been formed into a number of standardized tests to determine if adult learners can perform the competencies at deferent levels of ability. Because the test items are based on the competencies identified earlier, the CASAS tests are referred to as competency-based tests. [8]

In this regard, it should be noted that both the ABLE and TABE tests now provide competency-based information for interpreting individual items.

Curriculum-Based Assessment

Typically, in criterion-referenced or competency-based programs, developers first identify what the important objectives or competencies are that should be learned. Next, test items are developed to determine whether learners already possess the competencies or if instruction is needed to develop certain competencies. Then, various commercially available curriculum materials with a variety of learning exercises are identified that teach each of the competencies so that teachers can select the materials their learners need to master.

This approach, then, is a form of "teaching to the test," even though the exact contents of the assessment instruments may not appear in the curriculum to avoid directly teaching to the specific test items. The competency-based test is used, rather, to indicate the degree of transfer from the curriculum to the application of the new learning.

In curriculum-based assessment decisions are first made about what is important to be taught. Then a curriculum is developed, which may or may not be a formally, pre-developed series of learning experiences. Sometimes, very individualized content and learning activities are improvised by teachers and learners as a dynamic process. Finally, tests are constructed to "test to the teaching." Here the intent is to determine whether what is being taught is being learned and, if not, how instruction should be modified.[9]

In this case then, what is learned becomes the new competence gained in the program. The difference between the competency-based test and the curriculum-based test lies in the direction of test development. In the competency-based programs, the competencies are identified first and the curriculum is designed to help the learner achieve these specific competencies.

In the curriculum-based test, the learner's specific learning activities generate new competence that can then be certified through the development and administration of a curriculum-based test.

The idea of curriculum-based assessment arose from disappointment with the use of nationally standardized tests in which the contents and skills being assessed did not match precisely what was being taught in the schools.[10] This results in part from the requirement that, to market a test nationally, test developers cannot tie the test too closely to any particular curriculum. Further, they assess learning that takes place in both school and out-of-school experiences. As a consequence, the tests are generally not sensitive to the specific content (concepts; vocabulary; skills) that is being taught in a particular curriculum.

To appear to be related to all curricula, tests frequently use words that appear precise, but are not. For instance, assessing "Vocabulary Skills," as though "vocabulary" is a generalizable "skill," which it is not, instead of specific knowledge, which it is. In general, "skills"-oriented terminology is used to suggest that "process" ability and not content knowledge is being assessed. But this ignores that fact that all "process" requires some content on which to operate.

For adult basic education programs, in which there is generally precious little time for adults to participate, the "skills" focus is recognized as not being sensitive to the particular content that is taught. To a large extent, that is why there is very little increase in the standardized test scores of most adult learners in the relatively brief time that they attend programs. The nationally standardized and normed tests are not sensitive enough to the specifics of what is being taught in the program. Among others reasons, this is why many programs are searching for alternatives to such standardized tests. There is a desire for more curriculum-based assessment so that learners' "true" gains can be detected. This is discussed further under the topic of alternative assessment in Chapter 4.

Chapter 3

Review of Tests for ABE and ESL Programs[a]

There are hundreds of standardized tests. Yet only a very few have been developed for use by ABE or ESL program providers.

This chapter provides reviews of eight standardized tests that are widely used by ABE and ESL programs. These tests were selected for review to include the most widely used group-administered, norm-referenced tests of adult basic skills (ABLE, TABE); the group-administered, competency-based tests of CASAS; tests for ESL assessment (ESLOA; BEST; CASAS/ESL); tests that are used by volunteer adult literacy groups for individual testing in tutor-tutee arrangements (ESLOA; READ); and the GED Official Practice Test for indicating readiness for taking the GED high school equivalency examinations.

The information reported here for each test includes: the full name, commonly used acronym, and dates of publication; purpose; source; costs; description of skills assessed, reliability, validity, and types of scores that can be reported; and general comments. Notable strengths and weaknesses are high-lighted.

Reliability and validity coefficients are referred to as "low" when they are between 0 and .49, as "moderate" when between .50 and .79, and as "high" when equal to or greater than .80. When tests have different "levels" that means there are different tests for learners of different skill levels. The proper use of the appropriate level of test provides a more reliable estimate of learners' skills.

Wise use of these tests requires background knowledge that is not provided in this document or in the manuals that accompany most of the tests. Appendix B provides resources for further reading of a professional nature in the areas of testing and measurement. Final decisions about the use of any test should be made only after examining it carefully, reading its manual(s), and trying it with some students similar to those with whom it will be used.

Unless otherwise mentioned, the tests are suited to group administration, and the student test booklets are re-usable. The costs reported are for small orders and are only approximate, prices change over time; institutional or bulk order discounts are available from some publishers. Allow plenty of time when ordering materials. Order fulfillment normally takes 2-5 weeks unless special shipment and payment is specified. Errors in fulfilling orders are not uncommon.

---

[a]This chapter was written by Dr. Gregg Jackson. He holds a Ph.D. in Educational Research, and has worked in educational research and evaluation for 17 years. He has served as a volunteer in the D.C. Adult Education Program and, since 1987, has been a consultant to the Association for Comuunity Based Education's (ACBE) Adult Literacy Project. The reviews of tests are abstracts from more extensive reviews of 64 standardized tests and assessment instruments in a report prepared by Jackson for ACBE. See Appendix B for information on how to obtain the extended review of tests.

Adult Basic Learning Examination (ABLE, 1967-86)

Purpose: To measure several basic education skills of adults.

Source: The Psychological Corporation, Order Service Center, P.O. Box 839954, San Antonio TX 78283-3954; (800) 228-0752.

Costs: Learner test booklets cost $1.44; answer sheets cost $.50.

Description: There are sections on vocabulary, reading comprehension, spelling, language, number operations, and quantitative problem solving. There are three levels of the test, corresponding to skills commonly taught in grades 1-4, 5-8, and 9-12. There are two equivalent forms at each level for pre-and post-testing. A brief locator test is available to match the learners' skill levels to the appropriate level of test.

Reliability, Validity, and Scores: Test-retest reliability is not reported. Internal reliability has been high. Validity analyses show moderate correlations with the Stanford Achievement Test. Scores can be reported as scale scores, percentiles, stanines, and grade equivalents. Item response data are also reported. The norm data are based on 4,000 adults in 41 states and are reported separately for ABE/GED students, prisoners, vocational/technical students (only at Level 3), and a combination of all.

Comments: This is a 1986 revision of a test that has been widely used to evaluate the outcomes of adult basic education. The revision appears to be very responsive to several criticisms of prior tests used in adult basic education programs. The content and tone are adult. The reading passages are mostly about common everyday matters, and the questions tap not only literal comprehension, but also higher forms of comprehension. The mathematics word problems are representative of those many people encounter in daily life.

Ten of the items in the reading comprehension section of Level 1 (Form E) cannot be answered correctly without background knowledge that a moderate portion of adult learners will not possess or they require predicting what an imaginary person did in a given situation, and there is no way to know for sure. The "correct answer" presumes the imaginary person will act in the rational, safe, or common manner, but people do not always do so.

The Level 3 math section includes only a few very simple algebra and geometry problems. Some learners who score high may find themselves required to take remedial math when enrolling in technical schools and colleges.

This reviewer has extensive substantial experience in administering the reading comprehension and problem solving sections to adult literacy students. The students do not appear offended or antagonized by the test, they apply themselves and try to do well, and often perform somewhat better than their instructors had expected.

Basic English Skills Test (BEST, 1981-87)

Purpose: To assess speaking, listening, reading, and writing skills of low proficiency non-native English speakers.

Source: Center for Applied Linguistics, 1118 22nd Street N.W., Washington DC 20037; (202) 429-9292.

Costs: For the oral interview section, the administrator's picture cue books to which the learners respond cost $11.00 and answer sheets cost $.25; for the literacy skills section, the not re-usable learner test booklets and scoring sheets (together) cost $2.25.

Description: There are two sections. The oral interview section has 50 items and yields five scores for listening comprehension, pronunciation, communication, fluency, and reading/writing. It asks several personal questions, and then asks questions and gives the learners directions to follow in response to photographs, signs, a map, and some money placed on the table. The questions ask what are the people in the pictures doing, where is a specified object (the learner is to point to it), and what does a given sign mean. A few reading and writing items are included. The literacy skills section assesses reading and writing more thoroughly. There is only one level of the test. A second equivalent form of the test was recently made available.

Reliability, Validity, and Scores: Test-retest reliability is not reported in the manual. Internal reliability has been moderately high for the listening, communication, and fluency scores, and high for the total of the oral interview section. There are limited validity data. Learners assigned to seven ESL instructional levels, by means other than the BEST, were administered the BEST; the mean score of learners was substantially higher at each successive level. Though the test was administered to 987 ESL learners during its refinement, no norm data are reported in the manual. The manual describes "Student Performance Levels" for various total scores, but the basis for the specified levels is not given.

Comments: This test is adult in content and tone. The first section must be administered individually and to do so is moderately complex. Proper administration will require prior training and practice. The administration is paced and takes about 10 to 20 minutes. Most of the scoring of the first section is done as it is administered, not later from a tape recording. This saves time, but it can be distracting to the learner and sometimes even to the administrator. The scoring is judgmental and moderately complex, but after careful training inter-rater reliability has been high. A review of the test in Reviews of English Language Proficiency Tests (see Appendix B) described it as exciting, innovative, and valid, but time-consuming to use and lacking justification for the scoring system.

CASAS Adult Life Skills - Reading (1984-89)

Purpose: To assess a learner's ability to apply basic reading skills to common everyday life situations.

Source: CASAS, 2725 Congress Street #1-M, San Diego, CA 92110; (619) 298-4681.

Costs: Special training by CASAS is required before using this test; write or call for fees and material costs. CASAS is a Developer/Demonstrator Project in the U.S. Department of Education's National Diffusion Network.

Description: There is just one section of the test. Several levels are available, AA, A, B, C, suitable, respectively, for developmentally disabled and normal beginning, intermediate, and moderately advanced adult education learners. Level C is substantially easier than the GED test. There are two equivalent forms for each level. All CASAS tests are prepared from the CASAS item bank that now has 4,000 items. The bank permits quick and relatively inexpensive construction of customized tests for given objectives and difficulty levels. There are ready-made mathematics and English listening tests available.

Reliability, Validity, and Scores: Test-retest reliability is not reported. Internal reliability has been high. The manual and other publications sent to this reviewer do not indicate studies to validate the test against other measures of life-skills reading (though a moderate correlation of .70 was found in unpublished data for the ABLE and the CASAS reading test, see Appendix A, Table A-1 of this report). Raw scores are converted to CASAS scale scores; percentiles or grade equivalents are not reported.Data are presented for average entry, exit, and gains in programs throughout California over several years. Tables in the manual also indicate the specific objective measured by each item in the instruments.

Comments: This test is also referred to as the CASAS Survey Achievement Test. It is used widely in California by state-funded ABE and ESL programs, and it is also used elsewhere.The instrument is adult in content and tone. Virtually all of the reading materials are things that most adults would find very useful in everyday living. The content, however, is exclusively life-skill oriented. There are not items that use the kinds of reading material commonly found in popular magazines, newspapers, and books. Most of the items only assess literal reading comprehension. Few require inferences or evaluation.

Though CASAS is described as a competency-based assessment system, this reading test is not suited to assessing specific competencies. That is because the specified competencies are broad in scope and seldom measured by more than two items. For instance, in Form 31 of Level A, the competency of "interpret food packaging labels" is assessed by just one item, and the competency of "identify the months of the year and the days of the week" is assessed by only two items.

CASAS Adult Life Skills - Listening (1984-87)

Purpose: To assess English listening comprehension in common everyday life situations.

Source: CASAS, 2725 Congress Street #1-M, San Diego, CA 92110; (619) 298-4681.

Costs: Special training by CASAS is required before using this test; write or call for fees and material costs. CASAS is a Developer/Demonstrator Project in the U.S. Department of Education's National Diffusion Network.

Description: There are three levels, corresponding approximately to beginning, intermediate, and advanced ESL. There are two equivalent forms at each level. A cassette tape recording gives directions or asks a question, and the learner responds by selecting one of three alternative illustrations or sentences in a booklet. At the lowest level an example is: "Look at the pictures and listen [There are pictures of : a) a sheet of paper, b) a pencil, and c) a book]. What is the correct answer - A, B, or C? Give me a pencil. Is the answer A, B, or C?" At the low level, most items require no reading by the learners except of the letters "A," "B," and "C" used to designate the three pictures. At the intermediate level about half the items require reading at about the third grade level. At the high level, most of the items require reading at about the fifth grade level.

Reliability, Validity, and Scores: Reliability data are not reported in the materials examined. However, the test has been constructed in the same manner as several other CASAS tests that have had high internal reliability. Validity data are not provided.

Comments: Validity may be questionable. Many of the items in the intermediate and high levels of the test require reading skills. It is likely that some learners who comprehend the spoken English directions and questions are unable to select the appropriate responses because of inadequate reading skills. This would be particularly true in ESL programs serving learners who are illiterate in their native language and those that focus exclusively on oral language instruction methods.

A commendable array of life-skills materials are included, and most people living in the United States would find it useful to master the listening comprehension that is measured by this test. The test is used widely in California, and is also used elsewhere.

This is one of the few tests of oral English skills that does not have to be administered to one learner at a time. But because it was designed for group administration, it only assesses passive or receptive, not interactive or conversational comprehension of oral English. It also does not assess the speaking of English. Some learners have comprehension skills substantially above their speaking skills.

English as a Second Language Oral Assessment (ESLOA, 1978-80)

Purpose: To efficiently measure the ability of non-native English speakers to understand and speak English.

Source: Literacy Volunteers of America, 5795 Widewaters Parkway, Syracuse NY 13214; (315) 445-8000.

Costs: The cue books cost $7.25; answer sheets cost $.04.

Description: The test is divided into four progressively more difficult levels. There is only one form of the test. The learner is judged as being at level 1, 2, 3, or 4, depending on how many levels he or she completes. At the first level, the student is shown drawings with three objects and asked questions like: "Where is the Box?" or "Which girl is walking?" The learner may respond orally or by pointing. At the second level, the learner is asked to answer simple questions and name illustrated objects. At the third level, the learner is shown drawings and asked questions such as: "What is he doing?" and "Where is she going?" The learner must respond orally, and is encouraged to use complete sentences. The learner is also orally given several sentences and asked to modify them in a specified manner, such as from statements to questions. At the fourth level, the learner is orally given sentences and asked to change them to different tenses, shown pictures and asked what is happening in them, and told of specific circumstances and asked what he or she would do in them. There also is an optional section that provides a simple means for judging spoken English in response to personal questions such as: "What television shows do you like? Why?"

Reliability, Validity, and Scores: The publisher does not have reliability or validity data. The cue book, which also serves as the manual, does not report any norm data. Lesson content is suggested for learners who score at each of the four specified levels.

Comments: This test is part of the materials prepared and distributed by Literacy Volunteers of America. Most items deal with commonly encountered objects and events, but few directly involve the activities that most occupy adults' lives - working, meal preparation, housekeeping, and child raising. The test focuses on beginning and intermediate English. People scoring at the highest level, Level 4, could easily have difficulty understanding and participating in conversational English.

The test must be administered individually. Administration is simple and is terminated when a learner misses more than a specified number of items on any of the four sections. There is no time limit; 10 to 20 minutes will usually be needed. Scoring is simple and quick.

GED Official Practice Tests (1987-88)

Purpose: To help learner's determine their readiness to take the GED tests.

Source: Prentice-Hall, 200 Old Tappan Road, Old Tappan NJ 07675; (800) 223-1360

Costs: Learner booklets cost $2.13; answer sheets cost $.25.

Description: There are five sub-tests. They cover writing, social studies, science, interpreting literature and the arts, and mathematics. The GED tests cover the same subjects, but are about twice as long as the practice tests. There is only one level of the practice tests, but there are two English forms for use in the U.S., one for use in Canada, and one form entirely in Spanish.

Reliability, Validity, and Scores: Test-retest reliability, using the two equivalent U.S. forms, has been high for each sub-test, when assessed with a large sample of high school seniors. Internal reliability, based on data from a sample of GED candidates was also high. The sub-test scores on the U.S. forms correlated moderately highly with the comparable GED test scores in a large sample of high school students. Validity coefficients for GED candidates are not reported. Raw scores are converted to the same standard scale scores as used for the GED tests. The manual also reports the subject area and cognitive skill covered by each multiple-choice item. This can be used to help diagnose particular weaknesses that a learner may have.

Comments: This test was developed by the same organization that prepares the GED tests, and in accordance with the same specifications used for those tests. The test is adult in content and tone. The orientation is generally middle class and academic, but that is appropriate since the same is true of the GED tests.

This is a good predictor of GED test performance, and probably the best available. But all tests have some measurement error. For a learner to be reasonably assured of passing the GED in a state that requires passing every sub-test, all his or her predictor sub-test scores should be at least 13 GED scale points above the minimum pass level. That requires getting about two-thirds of the items correct in each sub-test.

Though there is no sub-test that specifically assesses reading skills, this test requires much reading, with most of it at about the 11th grade level. The test also requires considerable application of critical thinking.

Scoring of the essay part of the writing sub-test is complex, requires prior training, and is time consuming. An explanation of the procedures and accompanying examples take 53 pages in the manual.

Reading Evaluation Adult Diagnosis (Revised) (READ, 1972-82)

Purpose: To assess learner's reading needs and progress.

Source: Literacy Volunteers of America, 5795 Widewaters Parkway, Syracuse NY 13214; (315) 445-8000.

Costs: The cue books cost $7.25. Answer sheets, suitable for two administrations to the same learner, cost $1.25.

Description: The test has three parts. The first part assesses sight word recognition - identifying words without the application of phonic analysis. The learner is shown lists of words and asked to read them aloud. The easiest list includes words like "he" and "big;" the most difficult list includes words like "family" and "arrive." The second part assesses word analysis - the application of phonics to unfamiliar words. Learners are asked to name the letters of the alphabet, pronounce consonants, and pronounce words that may be unfamiliar. The third part assesses reading or listening comprehension. The learner is asked to read aloud, and to listen to short passages and answer questions about them - who, what, where, and how? There are two approximately equivalent forms of Part 1 and Part 3 of the test; there is only one form of Part 2.

Reliability, Validity, and Scores: No data on reliability are reported in the cue book, which also serves as a manual, nor in the supplemental information requested from the publisher. No data on validity are reported in the cue book. Supplemental information sent by the publisher indicates that a prior version of this test, prepared by a different author, correlated moderately with the reading scores from the Adult Basic Learning Examination (ABLE). That does not indicate the validity of the current version. No norm data are reported. Implications for instruction are provided with each section of the test.

Comments: This test is part of the materials prepared and distributed by Literacy Volunteers of America. It is intended to be used for diagnosis and monitoring. The reading difficulty ranges up to only about grade 5. The short reading passages are generally adult in orientation, but they seem bland to this reviewer and may not be of high interest to many low-income adults.

The test must be administered individually. The instructions are moderately complex, sometimes awkward to comply with, and occasionally incomplete. The complexity is caused by the variety of different types of items, each with its own instructions; dividing instructions for a given exercise among non-contiguous pages; interspersing pre-test and post-test items in the display materials; and specifying various skip patterns depending on the learner's performance. There is no time limit and no indication of how long the test normally takes to administer. Manual scoring is moderately complex, but takes only a few minutes for each student.

Tests of Adult Basic Education - Forms 5 and 6 (TABE, 1957-87)

Purpose: To measure reading, writing, and mathematics achievement.

Source: Publisher's Test Service, CTB/McGraw-Hill, 2500 Garden Road, Monterey CA 93940; (800) 538-9547.

Costs: The learner test booklets cost $1.62; answer sheets cost $.43.

Description: There are seven sections measuring vocabulary, reading comprehension, language mechanics, language expression, spelling, mathematical calculation, and mathematical concepts/application. There are four levels corresponding in difficulty to grades 2-4, 4-6, 6-8, and 8-12. A locator test is available for matching learner skill levels to test levels. There are two equivalent forms at each level.

Reliability, Validity, and Scores: Test-retest reliability is not reported in the manuals. Internal reliability has been high. Limited validity data are reported in the manuals. The scores on the TABE have correlated moderately with comparable scores on the GED. Scores can be reported as scale scores, percentiles, stanines, and grade equivalents. The norm data are based on 6,300 learners in 223 institutions across the country. Norms are reported separately for adult basic education learners, adult offenders, juvenile offenders, and vocational/technical school enrollees. Data in the Norms Book also permit prediction of GED scores, but should be treated as rough estimates because of the moderate correlations between the TABE scores and the GED scores. The Test Coordinator's Handbook reports the knowledge and type of cognitive skill covered by each test item.

Comments: The TABE is one of the most widely used tests in adult basic education programs. It was thoroughly revised in 1986. All the items are new, the range of skill levels that can be assessed has been extended, and the specific skills that are measured are more finely divided and identified.

However, the lowest level of the test will be daunting and frustrating for most students with less than grade 3.0 skills. For instance, the first reading exercise uses a 150-word passage.Though the items are adult in content, they seem to this reviewer distinctly middle class and academic in orientation. Only a modest portion of them are about everyday events in the lives of low-income adults. For instance, in the grade 4-6 level booklet (Form 5M), only two of the eight reading passages are about experiences common to such learners. Of the 40 items on math concepts and application there is only one item on calculating the correct change for a given transaction, no item on the savings from bulk purchases, and no item on the total cost of a purchase with installment plan financing charges. The language sections are notable for focusing on paragraph construction as well as sentence structure.

This test assesses an unusually broad range of skills. Therefore, giving the full TABE takes about 4.5 hours. For this reason, many programs use only one or two sections for pre- and post-testing.

Chapter 4

Special Topics in the Use of Standardized Tests

This section responds to questions about the uses of standardized tests and alternative assessment methods that policymakers, administrators, teachers, and evaluators have raised from time to time:

> What to do about "negative gain" scores, that is, when students do poorer at the end of the program than they did at the beginning?
>
> What is the difference between "general" and "specific" literacy and when should programs assess each?
>
> What is "item response theory" and what does it imply for testing in ABE and ESL programs.
>
> What is predictive validity and what does it have to do with assessment in ABE and ESL programs?
>
> What are some special problems in testing in ESL programs?
>
> What are "alternative assessment" methods and what are their advantages and disadvantages?
>
> What kind of assessment system can be developed to meet instructional purposes and State and federal requirements for accountability?

"NEGATIVE GAIN"

In ABE or ESL programs it is not unusual to find that 10-20 percent of learners score poorer on the post-test than they do on the pre-test. Therefore, when the post-test score is subtracted from the pre-test score to calculate the gain score, the gain is a negative number.[7,11]

It is possible (though not very probable, perhaps) that negative gain may occur because learners on the pre-test do not work at any given item too long, because they think they cannot perform the test task, and so they simply guess at all the items. On the post-test they spend more time on each item because they have new competence and think they should not guess but try to actually comprehend and perform each item. This could lead to more accurate, but fewer test items being completed at the post-test, and hence a negative gain score.

Generally, however, negative gain reflects guessing or other regression effects. In this case, guessing on the pre-test is better than guessing on the post-test and this leads to negative gain. This can be reduced by using tests that require constructed responses, or that offer many alternatives for multiple choice tests. The latter reduces the effects of guessing. In one study where tests with very low probability for guessing were introduced, negative gain was reduced from 30 percent to 6 percent.[5]

For those programs in which tests with higher potential for negative gain exists, and this includes all multiple choice tests, frequency distributions showing numbers and percentages of learners making various amounts of negative and zero gain should be included. This permits evaluators to gauge the amount of regression occuring in the program. Simply showing average pre-and post-test scores that includes the zero and negative gains obscures this valuable information and produces inaccurate indications of lower improvement in the program than actually occurs.

"GENERAL" AND "SPECIFIC" LITERACY

Learner-centered literacy instruction in which the functional context of the learner dictates the curriculum differs from literacy education based on the idea that adult basic education should replicate the school grades and eventually lead to a high school equivalency certificate. Literacy education aimed at giving the adult learner the same kinds of knowledge and information processing abilities as possessed by typical high school graduates is known as "general" literacy.

Literacy education aimed at providing adult learners with some particular, more circumscribed body of knowledge and information processing abilities, such as those involved in a particular line of work (e.g., automobile mechanic), life role (e.g., parent) or life activity (e.g., reading tax manuals) is known as "specific" literacy.

For many reasons, adult learners do not always have a lot of time to spend in a basic skills program. For instance, if they are unemployed and need to learn a job quickly, then time in a general literacy program that aims to recapitulate the public school curriculum will prolong the adult's entry into job training and hence into gainful employment. Furthermore, evidence suggests that "general" literacy education does not transfer much to improve "specific" literacy in the relatively brief (25,50,100) hours of education that adult learners will choose to attend. However, "specific" literacy training may produce as much improvement in "general" literacy as do typical "general" literacy programs.[5,12]

For these reasons, "workplace literacy" programs integrate basic skills training with job knowledge and skills development. For instance, a person desiring to learn to be an automobile mechanic is given reading, writing, and mathematics education using automobile mechanics training textbooks or technical manuals and performing functionally relevant, literacy task performance.

Following similar reasoning, if learners wish to read books to their children, literacy providers can teach "specific" literacy by teaching learners about children's books, how to read and interpret them with their children, and so forth. Or, adults desiring to read a tax manual can be taught literacy using a tax manual and special materials to develop "specific" ability in reading tax manuals.

A very large amount of materials and procedures exist for teaching English for Specific Purposes (ESP) in English as a Foreign Language or in English as a Second Language (ESL) programs. Such ESL programs are sometimes known as VESL-Vocational English as a Second Language- programs.

In all these specific literacy or language programs, assessment instruments can be developed that are curriculum-based, as discussed above. These "specific literacy tests" will be most sensitive to the adult learners' goals and gains. Programs can also use "general literacy" tests to indicate the degree of generalizability that occurs in the "specific" literacy program.

ITEM RESPONSE THEORY (IRT)

With the growth in use of tests such as the Comprehensive Adult Student Assessment System (CASAS)[8] and the National Assessment of Educational Progress (NAEP) young adult literacy profile,[13] more ABE and ESL program providers are reading about *item response theory.*

The CASAS and NAEP (as well as the TABE, ABLE, Degrees of Reading Power, and several other tests) have been developed using newer psychometric methods based on item response theory. In general, IRT is a method for scaling individual test items for difficulty in such a way that the item has a known probability of being correctly completed by an adult of a given ability level.[14] For instance, on the CASAS scale, an adult learner with an ability score of 215 has a fifty percent chance of passing all items that are in the item bank that are also scaled at 215. For items rated below 215, the learner has a greater than fifty percent chance of getting the items correct, and with items above 215 the learner has less than a fifty percent chance of getting the items correct.

If a program has a test item bank of several thousand items that are all on the same IRT scale, it is possible to administer a relatively small sample of the items in a test and from this small sample of items, know the probability that the learner can perform each of the other items in the bank. Obviously this is useful for diagnosing particular competencies that a learner may need to develop further.

Traditionally developed tests do not provide probability of performance estimates for items not in the test. Furthermore, traditionally developed, norm-referenced tests have to be renormed everytime the items in the test are changed. But with an IRT-based test, items from a bank can be reconfigured into different forms of tests without having to renorm the test. This means that it is easier for programs to tailor tests for their particular curriculum and for learner needs.

In particular, IRT is useful for developing multiple forms of tests that are suitable for a restricted range of ability. This permits more reliable estimation of ability for learners within the range being studied.

Though the power of IRT will ensure that most future test development will utilize this psychometric technology, it should be noted that there is nothing in the IRT that ensures the *validity* of the tests. Validity refers to whether or not a test actually measures what it purports to measure, and nothing else.

But absolute validity is a very difficult thing to achieve. All paragraph reading comprehension tests, for instance, measure not only skill in decoding printed language and performing tasks such as identifying the main idea, but also a learner's background knowledge related to what is being read. This is true regardless of whether the tests are developed using traditional or item response theory psychometrics.

PREDICTIVE VALIDITY

In the discussion of Item Response Theory, *validity* was defined as referring to whether or not a test measures what it purports to measure and only that.

There is, however, another type of validity that is assuming greater importance in ABE and ESL. This type of validity is called *predictive validity* . Predictive validity refers to how valid or accurate a test is for predicting some future behavior of learners. It is growing in importance as such federally mandated programs as the Job Training Partnership Act (JTPA), the Job-Oriented Basic Skills (JOBS) workfare/welfare program, and workforce literacy programs focus on identifying participants whose basic skills are judged to be too low for employment. Under such programs, adults identified as "functionally illiterate" may be denied job training because of their low levels of basic skills. They may be required, instead, to participate in basic skills courses to qualify for job training or to continue to receive their welfare benefits, or both.

Predictive validity is also important in pre-GED testing to determine whether learners qualify to attempt the GED high school equivalency examination. For instance, the CASAS scales suggest that learners with scores of 224 or below are functioning below a high school level, while those with scores at or above 225 can profit from instruction in GED preparation.[15]

The Official GED Practice Tests are used "...to provide *general* indications of readiness to take the full-length GED Tests."[16]

All uses of basic skills tests to indicate "readiness," ability to "profit from instruction" and that prevent learners from entering into some desired job or job training program are predicting that learners who score below a certain level on the basic skills test will not be successful in the future activity for which the basic skills test serves as a screen. The question for predictive validity is, does the test score criterion accurately (that is, validly) predict who will and will not be able to perform satisfactorily in the job, job training, or GED test-taking situation?

In studies of the predictive validity of the most widely used basic skills test, the Armed Forces Qualification Test (AFQT), it was found that of those that military selection policies had predicted to fail in job training and on the job, eight out of ten actually performed satisfactorily.[17] These data, from an organization that has studied this type of assessment for seventy years at a cost of at least $100 million, should caution the "gatekeeping" use of basic skills tests in workfare/welfare, workplace literacy, and JTPA programs.

No major gatekeeping decision should be based solely on the results of a single standardized test score. Adult education providers should use interviews, past

employment experiences, and work sample procedures to counsel learners about their probabilities of success in future activities beyond the boundaries of the basic skills program.

There are well-established laws, and many precedent-setting legal cases to establish a basis for adult learners to challenge test use that adversely impacts them by delaying or preventing access to gainful employment.[18] To date, no studies have been found of the predictive validity of standardized tests used in workfare/welfare basic skills programs, workplace literacy programs or GED preparation programs.

ENGLISH AS A SECOND LANGUAGE

A growing share of adult basic education is concerned with English as a Second Language programs. In 1985-86, ESL participants made-up 57 percent of ABE students nationally.[19] In California, ESL learners make-up close to 80 percent of participants in ABE.[20]

Using standardized tests with ESL learners incorporates all of the problems discussed earlier in this report. Additionally, however, special difficulties are encountered because of the wide differences in the language, cultural, and educational backgrounds of the ESL learners.

For instance, many ESL learners come from groups for which there is no written language (e.g., Hmong, Mien) and so it cannot be assumed that they have general, "world" knowledge of the forms and uses of written language.[21] Others, however, may be highly educated and literate in their native language, but simply unable to speak and comprehend English. Given this large range of differences among ESL learners, there is a need to determine, through interviews with learners or their associates, the non-English language education and literacy status of ESL learners prior to administering assessment instruments.

The major difference between ABE and ESL students, of course, is their knowledge of the English language. Most adults, even the highly literate and educated, are reticent about speaking a foreign language. ESL learners are no different from other adults in this regard. Hence, it is necessary to have a period of adjustment during which learners can develop confidence before proceeding with a formal assessment using standardized tests that require learners to speak. This is similar to the need for a "warm-up" period discussed above.

Because speech disappears as it is produced, the evaluation of English speaking, comprehension, and communicative functioning ability (e.g, knowledge of forms of speech for particular occasions) in a dynamic interaction is difficult. This may lead to test situations in which the types of tasks called for are designed to permit special judgments for ease of scoring to be arrived at, but which also appear "unreal" to both teachers and learners. For instance, standardized tests may not permit normal conversational patterns, questioning of meanings by learners, and sharing of information to accomplish a real-life task.[22] This may lead to an underestimate of the learner's communicative competence.

Generally, in testing in ESL programs, as in other ABE programs, it may be desirable to separate testing for program accountability from testing for instructional decision making.

ALTERNATIVE ASSESSMENT METHODS

Problems involved in obtaining valid measures of learners' development in adult literacy programs have stimulated a growing interest in alternatives to standardized tests for assessing learner's progress in instructional programs.

The September 1989 issue of Information Update, the newsletter of the Literacy Assistance Center, Inc. in New York focuses on alternative assessment methods. The issue provides a good example of the types of problems that program providers experience with standardized tests, and presents a rationale for the need for improved assessment methods.

The major problem addressed by the alternative assessment movement is similar to that discussed under curriculum-based assessment, namely the incongruence between what programs teach, what learner's learn, and what the nationally standardized tests assess.

Many of the programs that are experimenting with alternative assessment methods do not follow a prescribed curriculum. Rather, they follow an approach in which a learner's expressed needs form the basis for instruction. This approach is frequently called a *learner-centered* or *participatory* approach, because the learner participates in determining the instruction.[23]

Alternatives to nationally standardized testing include intake and progress interviews that record such information as the type of reading the learner does, how much reading in different domains (job, home, community) is accomplished, self-evaluations of reading ability, and judgments of abilities by teachers in staff meetings. The California Adult Learner Progress Evaluation Process (CALPEP) illustrates the interview approach to assessment.[24]

A second method of alternative assessment is portfolio development and evaluation.[25] This is a method similar to that followed by artists, designers, models, writers and others in creative fields of endeavor. Using this method, learners develop portfolios of their work in reading, writing, and mathematics, including both in-class and out-of-class work. Peers, teachers, and learners meet periodically to discuss the learner's work and how it is progressing.

Through these meetings, learners' progress is assessed in areas such as *metacognitive* (thinking about, evaluating, and planning their work), *cognitive* (vocabulary, concept knowledge, and reasoning processes typical of an area chosen by the learner; knowledge of the functions and structure of various types of texts -notes, letters, reports from school, work materials, etc.), and *affective* (self-understanding and esteem, value of literacy for self, children, and others).

Sometimes direct indicators of competence and its change are obtained by having learners perform, much as a performing artist would. For instance, in a reading program the performance might consist of *reading aloud*.[9,26] As the learner performs, the teacher may record the oral reading and then later listen to the recording with the learner. Together they evaluate the reading performance for progress in pronunciation, accuracy of word identification, inflection cues to comprehension, and other information identified in participation with the learner.

Assessing Alternative Assessment

There can be no doubting that the alternative assessment methods provide new information about adult learners in ABE and ESL programs. Much of this information reflects newer concepts about literacy and other abilities from contemporary cognitive science.

Alternative assessment methods relate very much to the teaching and learning process as it takes place in the classroom in interactions among teachers, learners, peers and the various materials they use and tasks they perform. In general, the richer the descriptive information about these interactions and processes, the more valid will be the understanding of particular programs by both internal (administrators; teachers; learners) and external (local; state; federal) evaluators.

However, while these alternative methods are invaluable for their contributions to learner progress, there are limitations to the exclusive use of such techniques for learner and program evaluation, as those developing these new assessment methods acknowledge.[27]

One of the problems identified by alternative assessment providers is the fact that, although standardized, nationally normed tests fail to match program content, administrators, teachers, and millions of other adults can and do perform very well on any or all of the dozens of standardized tests of reading, writing, and arithmetic that are the subject of criticism. The question is raised, therefore, of whether or not adult learners in ABE and ESL programs are being directed to less demanding levels of achievement if they are not evaluated using standardized tests.

It has also been noted that standardized tests

> "...are an integral part of the fabric of our lives. One has to take tests to get into college, to enter the military and to obtain civil service employment, to mention just a few. While such tests should certainly not be the measure of individual student progress in the adult literacy classroom, we ought not ignore the value for students of being familiar with them and being able to use them to their own advantage."[27]

A problem with the sole reliance on alternative assessment methods for program evaluation for public accountability is that nonstandardized methods make it difficult to compare across programs. One goal of the federal guidance on standardized tests is to make it possible for outside evaluators to know how well one program or group of programs is promoting learning compared to other programs. The intent is not to evaluate

individual learner growth and development. That is why the federal comments suggest that not all individuals must be assessed using standardized tests. Rather, a representative sampling scheme may be used to represent program and not individual outcomes.

ASSESSING FOR INSTRUCTION AND ACCOUNTABLITY

Many of the problems with standardized testing experienced by programs are due to the attempt to use one test for both program accountability and instructional decision making. For instance, using the TABE for pre and post-testing to report gains in general literacy to state and federal administrators is a program accountability function of the tests.

But using the TABE to assess learning in a specific literacy program, in which learners may choose to read and study a technical manual is an inappropriate use of the test for assessing either instructional needs or progress. In this case, an alternative assessment method is needed, perhaps one in which learners' needs are determined by interviews that include trial readings of technical manual passages. Then, progress checks using reading aloud and question/discussion periods for checking comprehension might be used to indicate learning in the program.

In one military project, a specific job-related literacy program was developed that used three types of testing.[5] Pre and post-module testing was used in a competency-based, criterion-referenced, testing/teaching curriculum strand. The module tests provided curriculum-based indicators of both instructional needs and progress.

A second testing method was developed in which job-related reading tasks from across six career fields were selected and included in job-related reading task tests. These tests were used as pre and post-program measures of generalizable growth in work-related (though not job-specific) types of reading skills. They were then normed in grade levels because the military management preferred to indicate program growth in grade levels.

Finally, a nationally standardized and normed test was administered pre and post-course to indicate growth in general literacy in grade level units.

As might be expected, in this program, the most learning was indicated by the pre and post-module tests, the next largest increase was in the pre and post-course, work-related tests, and the least increase was in the general literacy tests.

In general, multiple assessments can contribute multiple types of information. Nationally standardized tests, properly administrated, can provide information about broad growth in literacy or mathematics skills.

But this growth will typically not exceed one or two "years" in 25, 50 or 100 hours (and this must be obtained with regard to the problems of warm-up and regression discussed earlier). This information can be used for cross-program evaluations of broad ability development.

For instructional decision making, assessment more closely coupled to the curriculum provides the best indicator of what is being achieved by learners in the program. In general, the two important questions here are, "What do learners want to learn?" and "Are they learning it?"

In some competency-based programs, such as those using the CASAS, the testing is designed so that it can be used for both accountability and instructional decision making. By continually enlarging the number of items in the item pool, psychometrically sound tests can be tailor-made for a large number of learner-determined domains of instruction. These tests can then be interpreted in terms of the particular general ability scale involved.

Footnotes

[1] Public Law 100-297, Title III, Part A, Subpart 5, section 352: Evaluation.

[2] Federal Register, August 18, 1989, p. 34435.

[3] T. Sticht (1982). Evaluation of the "reading potential" concept for marginally literate adults. Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics).

[4] Regression to the mean also ocurs whenever a high scoring group has been separated from the total group and retested later on. In this case, the average score of the high scoring group will tend to decrease as it regresses to the mean of the total group.

[5] T. Sticht (1975). A program of Army functional job reading training: Development, implementation, and delivery systems (Final Report). HumRRO-FR-WD-(CA)-75-7. Alexandria, VA: Human Resources Research Organization.

[6] R. Glaser & D. Klaus (1962). Proficiency measurement: Assessing human performance. In: R. Gagne (Ed.), Psychological principles in system development. New York: Holt, Rinehart, and Winston.

[7] R. Taggart (1985, March). The Comprehensive Competencies Program: A summary of 1984 results. Washington, DC: Remediation and Training Institute.

[8] J. Davis, P. Williams, R. Stiles, J. Wise & P. Rickard (1984, April). CASAS: An effective measurement system for life skills. New Orlieans, LA: Paper presented at a meeting of the National Council on Measurement in Education.

[9] R. Bean, A. Byra, R. Johnson, & S. Lane (1988, July). Using curriculum-based measures to identify and monitor progress in an adult basic education program. Final Report. Pittsburgh, PA: University of Pittsburgh, Institute for Practice and Research in Education.

[10] L. Fuchs & S. Deno (1981). The relationship between curriculum-based mastery measures and standardized achievement tests in reading. (Research Report No. 57). Minneapolis, MN: University of Minnesota, Institute for Research on Learning Disabilities.

[11] J. Caylor & T. Sticht (1974, April). The problem of negative gain scores in the evaluation of reading programs. Chicago, IL: Paper presented at the annual meeting of the American Educational Research Association.

[12] T. Sticht (1988). Adult literacy education. In: E. Rothkopf (Ed.), Review of research in education. Volume 15. Washington, DC: American Educational Research Association.

[13] I. Kirsch & A. Jungeblut (1986). Literacy: Profiles of America's young adults. Princeton, NJ: Educational Testing Service.

[14] More can be learned about Item Response Theory (IRT) in a text and computer assisted instruction program: F. Baker (1985). The BASICS of item response theory. Portsmouth, NH: Heinemann Educational Books.

[15] The CASAS System (1989, November). GAIN Appraisal Program. III. Third Report. San Diego, CA: San Diego Community College District Foundation, Comprehensive Adult Student Assessment System.

[16] American Council on Education (1989). The Official GED practice Tests. Washington, DC: American Council on Education.

[17] T. Sticht, W. Armstrong, D. Hickey & J. Caylor (1987). Cast-off youth: Policy and training methods from the military experience. New York: Praeger.

[18] B. Gifford (Ed.) (1989). Test policy and the politics of opportunity allocation: The workplace and the law. Boston, MA: Kluver Academic Publishers.

[19] R. Pugsley (1987). National data update. Washington, DC: U.S. Department of Education, Paper presented at the Annual Conference, State Directors of Adult Education.

[20] D. Dixon, M. Vargo & D. Campbell (1987, July). Illiteracy in California: Needs, services & prospects. Palo Alto, CA: SRA Associates.

[21] K. Savage (undated, circa 1983). Teaching strategies for developing literacy skills in non-native speakers of English. Washington, DC: The National Institute of Education.

[22] P. Tirone (unpublished, circa 1988). Teaching and testing - can we keep our balance. New York: Literacy Assistance Center, Inc.

[23] S. Lytle, A. Belzer, K. Schultz & M. Vannozzi (1989). Learner-centered literacy assessment: An evolving process. In: A. Fingeret & P. Jurmo (Eds.), Participatory literacy education. San Francisco, CA: Jossey-Bass.

[24] R. Solorzano (1989, February). Analysis of learner progress from the first reporting cycle of the CALPEP field test. A report to the California State Librarian. Pasadena, CA: Educational Testing Service.

[25] See articles by M. Wolfe and S. Hill in the September 1989 special issue of Information Update published by the Literacy Assistance Center, Inc. of New York city. For an earlier application of portfolio-type assessment applied to adult education see R. Nickse (1980). Assessing life-skills competence. Belmont, CA: Pitman Learning, Inc.

[26] S. Hill (1989, September). Alternative assessment strategies: Some suggestions for teachers. Information Update, 6, pp. 7,9.

[27] M. Dick (1989, September). From the editor. Information Update, 6, pp. 1-2.

[28] R. Morris, L. Strumpf & S. Curnan (1988, May). Using basic skills testing to improve the effectiveness of remediation in employment and training programs for youth. Washington, DC: National Commission for Employment Policy.

[30] T. Sticht (1985). Understanding readers and their uses of texts. In: T. Duffy & R. Waller (eds.), Designing usable texts. New York: Academic Press.

[31] W. Grimes & W. Armstrong (unpublished, 1988-89). Test score data for the ABLE and CASAS survey of achievement tests. San Diego, CA: San Diego Community College District, Division of Continuing Education.

[32] Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs and Logistics) (1982, March). Profile of American youth. Washington, DC.

[33] B. Waters, J. Barnes, P. Foley, S. Steinhaus & D. Brown (1988, October). Estimating the reading skills of military applicants: Development of an ASVAB to RGL conversion table. Final Report 88-82 (HumRRO FR-PRD-88-82). Alexandria, VA: Human Resources Research Organization.

Appendix A

Correspondences Among Frequently Used Tests

A recent survey of 150 basic skills programs funded by the Job Training Partnership Act (JTPA) found that the four most frequently used standardized tests were the Tests of Adult Basic Education (TABE), the California Achievement Test (CAT), the Wide-Range Achievement Test(WRAT), and the Adult Basic Learning Examination (ABLE).[28]

An earlier national survey of adult basic education (ABE) programs, including civilian and military, school and community based programs found that the TABE was the most widely used standardized test, and the ABLE was used almost exclusively in community based programs.[29]

Of the more recently developed adult literacy assessment instruments that use item response theory scales exclusively, the Comprehensive Adult Student Assessment System (CASAS) is being used in hundreds of programs in 25 states.

The most widely used test of adult basic skills is the Armed Forces Qualification Test (AFQT), which, in both high school vocational counseling forms and military application forms is taken by over a million and a half youth and young adults yearly.

Given the widespread use of the TABE, ABLE, and CASAS tests, it is useful to have an idea about what scores on one of the tests means in terms of scores on the others.

Unfortunately, no study has been found that correlates these tests on a national sample. There are, however, a number of smaller-scale studies that permit the formulation of very rough correspondences. In an unpublished study by the U. S. Army the then current forms of the mid-level ABLE and TABE reading tests were administered to a sample of several thousand military personnel. This permits one estimation of the correspondence among those tests. [30]

In another unpublished study, correlations among the 1986 edition of the mid-level (level 2) ABLE and CASAS (levels B and C) tests were obtained for some 600 participants in the California Greater Avenues for Independence (GAIN) welfare program.[31]

In a published study using a nationally representative sample of young adults, the Department of Defense presented average Armed Forces Qualification Test (AFQT) scores for several different subpopulations, along with the average estimated ABLE reading test scores for these same subpopulations. This permits the derivation of AFQT scores from ABLE scores.[32]

Using the data from these various published and unpublished studies, it is possible to estimate ABLE scores from TABE scores, CASAS scores from ABLE scores, and AFQT percentile scores from ABLE scores. These correspondences are shown in Table A-1.

Table A-1
Correspondences Among Widely Used Tests of Adult Basic Skills

TABE[a]	ABLE[a]	CASAS[b]	AFQT[c]	NAEP SCALE
3.0	3.5	213	--	
4.0	4.3	215	--	Rudimentary/
5.0	5.2	217	05	Basic
6.0	6.0	220	14	
7.0	6.8	222	23	
8.0	7.7	224	33	Intermediate
9.0	8.5	226	41	
10.0	9.4	229	50	
11.0	10.3	231	60	Adept
12.0	11.0	233	68	

[a] Reading grade level scores. [b] CASAS scale scores. [c] Percentile scores
NOTE: The AFQT correspondence to ABLE is based on an earlier form of that test. However, more recent studies (33) confirm the correlation of new forms of both TABE and ABLE relationships at about .85 with the new and old versions of the AFQT. The NAEP also includes a scale category above "Adept" called "Advanced."

Caveats. As noted above, Table A-1 is presented to permit program operators to have a very rough idea of the correspondence that may exist among these widely used standardized tests of basic skills. The table is based on the logic that if A = B, and B = C, then A = C. The TABE is A, The ABLE is B, and CASAS is C. Of course, this is not nearly as desirable as direct studies to intercorrelate these tests. Furthermore, in addition to the possibility of error in estimation resulting from the failure to have TABE and CASAS correlations, and therefore having to estimate the relationship of the TABE to CASAS via the ABLE, it should be recalled that the TABE and ABLE correspondence is based on earlier versions of these tests, not the current editions while the ABLE and CASAS correspondence is based on current editions.

In the Grimes & Armstrong[31] study CASAS tests levels A and B were approximately matched to learner's entering reading levels, only Level 2 (the mid-level) of the ABLE test was administered. This makes estimates at both the high and low ends of the scale less reliable. It should also be noted that the CASAS scores in Table A-1 were estimated from ABLE scores. It is instructive to note that, if the reverse procedure is used, that is, if ABLE scores are estimated from CASAS scores, that does not give the same numbers as obtained in estimating CASAS from ABLE scores as in Table A-1. Thus, if a program operator is using the CASAS and wants to estimate ABLE scores, the following table should be used (r=.71):

Table A-2
Using CASAS Scores to Estimate ABLE Scores

If CASAS score is 200	then ABLE score is 3.9
215	6.6
225	8.5
230	9.4

In an unpublished study by a CASAS affiliate in Missouri, some 460 TABE scores were obtained along with CASAS scores on the ECS (Employability Competency System), a test similar to the the test used above, with the ABLE. Table A-3 shows correspondences among selected CASAS and TABE scores (r=.77).

Table A-3
Using CASAS Scores to Estimate TABE Scores

If CASAS score is 200	then TABE score is 4.2
215	7.0
225	8.8
230	9.8

In interpreting the data of Table A-1, it is interesting to note that the 1985 assessment of young adult literacy by the National Assessment of Educational Progress (NAEP) estimated that there were about 4-5 percent of young adults (21-25 years old) reading below the level of the typical fourth grade student (a student in the range from 4.0-4.9).[13] In Table A-1, the Department of Defense data show that a score of 5.0 on the TABE or 5.2 on the ABLE corresponds to about the 5th percentile on the AFQT, suggesting that about four percent of young adults 18-23 years of age score in the fourth grade or below levels. This is a surprisingly close estimate to the NAEP data.

The CASAS project interprets the CASAS scale scores and suggests that below a score of 200 (ABLE 3.9, Table A-2; TABE 4.2, Table A-3) learners have difficulty with the basic literacy and computational skills needed to function adequately in employment and in the community. Given the data of Table A-1 showing that young adults scoring below the score of 200 on the CASAS score at or below the 3rd grade level on the TABE and ABLE, and below the 5th percentile on the AFQT, the CASAS interpretation seems appropriate, and perhaps a bit conservative.

It is also consistent with the categories of reading proficiency given by the National Assessment of Educational Progress (NAEP) in its study of young adult literacy.[13, pp. v-1-v-3] Table A-1 shows the NAEP categories in relation to the other test data. This correspondence is based on data showing that some 16 percent of young adults scored below the Rudimentary and Basic levels on the NAEP, and this is close to the14 percent scoring at or below the 6th grade reading levels on the TABE and ABLE and the CASAS score of 220.

The latter is, in turn, near the cutoff of 215 (ABLE 6.6 grade; TABE 7.0) below which the CASAS defines as indicating those adult learners having low literacy skills and who "are functioning below a 7th grade level."[15, p. 9] Again, this interpretation seems to agree quite well with the TABE and ABLE data.

The CASAS scale scores between 215-224 are interpreted by the CASAS as indicative of lower than high school entry functioning, and indicate the need for pre-GED education. For scores of 225 (ABLE 8.5, Table A-2; TABE 8.8, Table A-3) and above, learners are said to be able to function at a high school entry level and can profit from instruction at the high school level. These interpretations, too, seem reasonable given the finding that CASAS scores of 226 or above correspond roughly to high school level reading in the 9th to 12th grade levels on the TABE and ABLE tests (Table A-1).

The NAEP category of **Intermediate** includes some 30 percent of young adults and this corresponds roughly to the 30 percent of young adults scoring between the 20th and 50th percentiles on the AFQT as indicated in Table A-1.

Given the margin for error in all these measurement instruments and systems, and the differences that result from changing which scores are used to estimate the others, the data of Tables A-1, A-2 and A-3 should be regarded as providing only rough correspondences. For instance, given their standard errors of measurement,TABE or ABLE scores of 6.0 might really be scores of 5.0 or 7.0. A CASAS score might vary from 5 to 10 points depending upon the particular form and the test population. The AFQT percentiles are subject to measurement error of plus or minus some 20 points, while the NAEP scale scores may also vary by several score values.

Perhaps, if the present data seem useful, a large scale study to more precisely establish these correspondences will be conducted. However, even if this were done, it should be recognized that a major point to be derived from this paper is that all three of these measures, and indeed of all tests following even the most advanced of psychometric methods are imprecise. And no new study will overcome this problem. There is no one test, or no one fixed score that indicates the "true" skill level of an individual or a group.

Because of the error in these types of standardized tests, no major gatekeeping decisions should be based solely on a single "cut" score on a single test. Rigid rules should not be established such as saying that all who score below a CASAS score of 225 or below an 8.9 grade level should be sent to basic skills education. Rather, there should always be multiple sources and types of information about people, including past histories of achievements, employment, informal samples of performance using basic skills, references and other types of information that can help in the decision making process.

Appendix B

Resource Materials

In addition to the references cited in the Footnotes of this report, readers interested in learning more about measurement and test theory and its applications will find these "classics" readable and useful:

A. Anastasi (1982). Psychological testing. 5th ed. New York: MacMillan Co.

L. Cronbach (1984). Essentials of psychological testing. 4th ed. New York: Harper and Row.

The Standards for Educational and Psychological Testing (1985) were developed by the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education to provide guidance on the proper development and use of standardized tests. Copies may be obtained from the American Psychological Association, Washington, DC.

A reference librarian can assist in locating further reviews of basic skills tests in such volumes as Reviews of English Language Proficiency Tests or the various editions of Buro's Mental Measurement Yearbook.

Until recently, there has been no comprehensive review of tests for ABE or ESL program providers. This has been remedied with the highly competent work by Dr. Gregg Jackson for the Association for Community Based Education (ACBE).

Entitled Measures of Adult Literacy Program Outcomes, this work reviews 64 instruments that are available for evaluating a variety of outcomes of adult education programs.

The reviews of tests in Chapter 3 of the present report were prepared by Jackson. For the most part, they are brief digests of the more extensive reviews in the ACBE report.The longer reviews incorporate more of Jackson's personal experiences in administering and interpreting several of the tests, and contain the types of comments that can be particularly useful for adult educators. The report may be obtained in April, 1990 from the

ERIC Clearinghouse on Tests, Measurements, and Evaluation
American Institutes for Research
3333 K St., NW, Suite 200
Washington, DC 20007
(202) 342-5060

Reviews of tests for the evaluation of programs for limited English proficient students, adults or children, are available from

Evaluation Assistance Center (East)
Georgetown University
1916 Wilson Boulevard, Suite 302
Arlington, VA 22201
(800) 626-5443

The Armed Forces Qualification Test (AFQT) is not available to the general education community. However, information about the AFQT and the Armed Services Vocational Aptitude Battery (ASVAB) and its use in civilian schools for vocational counseling is available from

Military Entrance Processing Command
2500 Greenbay Road
North Chicago, IL 60064

Information about the National Assessment of Educational Progress (NAEP) Adult Literacy Profile test is available from

National Assessment of Educational Progress (NAEP)
Adult Literacy Profile
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Information about the General Educational Development (GED) tests and predictor tests may be obtained from

GED Testing Service
American Council on Education
One Dupont Circle, N.W.
Washington, DC 20036

Policy issues related to the testing and assessment of adult literacy are addressed in a report entitled "Enhancing Adult Literacy: A Policy Guide" available from

The Council of State Policy & Planning Agencies
Hall of the States
400 North Capitol, Room 291
Washington, DC: 20001
(202) 624-5824

A very extensive bibliography on the politics of testing, new assessment methods, employment and the legal aspects of testing and other topics is available from

The National Commission on Testing and Public Policy
Boston College
McGuinn Hall 530
Chestnut Hill, MA 02167
(617) 552-4516

For materials that offer critical insights about educational and employment problems that may result from the inappropriate use of standardized tests contact

The National Center for Fair & Open Testing
342 Broadway
Cambridge, MA 02139-1802
(617) 864-4810

STAFF DEVELOPMENT MATERIALS

State adult education policymakers, administrators, teachers and others involved in ABE and ESL programs may wish to draw on information in this report as part of their staff development activities. For this reason, the following pages include a set of transparency masters that may be useful in developing presentations on standardized testing in ABE and ESL programs.

Testing and Assessment in ABE and ESL Programs

The Adult Education Act, Amendments of 1988, requires that:

States evaluate at least one-third of grant recipients.

Among other things, data from standardized tests are to be reported.

This presentation includes

- Definition of standardized test.
- Types of standardized tests.
- Things to avoid and to do to use standardized tests properly.
- Reviews of eight widely used standardized tests.
- Tables showing correspondences among widely used tests.
- Alternatives (or supplements) to standardized tests for evaluating learner progress and program outcomes.

Testing and Assessment in ABE and ESL Programs

STANDARDIZED TEST:	**A test administered under standard conditions so the scores reflect the skills being assessed and nothing else.**

TYPES OF STANDARDIZED TESTS

NORM-REFERENCED:	**A test in which a learner's score is compared to the scores of others who have taken the test. Example: Reading at a 7.0 grade level means the learner scored on the test like a typical child at the beginning of the 7th grade. The score does not tell how well a domain of skill has been learned. Items are chosen to differentiate among people.**
CRITERION-REFERENCED	**A test in which a learner's score is compared to an absolute standard, such as 80, 90, or 100 percent mastery of a domain of skill. Items are chosen on the basis of their importance, not how well they differentiate among people.**
COMPETENCY-BASED	**A test in which test items are made to measure stated competencies or objectives.**
CURRICULUM-BASED	**A test in which items or tasks are developed to determine if what is being taught is being learned.**

Testing and Assessment in ABE and ESL Programs

STANDARDIZED TEST: A test administered under standard conditions so the scores reflect the skills being assessed and nothing else.

AVOID	DO
VIOLATIONS OF STANDARD CONDITIONS	READ THE MANUAL!
IGNORING TIME LIMITS	USE STATED TIME LIMITS
USING WRONG LEVEL OF TEST	USE APPROPRIATE LEVEL
GIVING TESTS THE FIRST DAY OF CLASS WHEN LEARNERS MAY BE TIRED ANXIOUS, LOW IN TEST-TAKING SKILLS	BEFORE PRE-TESTING,LET LEARNERS ADJUST TO PROGRAM; PROVIDE ANXIETY REDUCTION EXPERIENCES; DEVELOP TEST-TAKING SKILLS

EIGHT TESTS WIDELY USED IN ABE AND ESL PROGRAMS

Test	Acronym	Purpose	Tests	Norms
Adult Basic Learning Examination	ABLE	Measure basic skills	Groups	Child & Adult
Basic English Skills Test	BEST	Measure English langauge Skills	Indvds.	Not Reported
CASAS Adult Life Skills-Reading	CASAS/ READ	Measure life skills in reading	Groups	Adult
CASAS Adult Life Skills-Listening	CASAS/ LISTEN	Measure life skills in ESL listening	Groups	Adult
English as a Second Language Oral Assessment	ESLOA	Measure English language skills	Indvds.	None Reported
GED Official Practice Tests	GED/ PRAC	Measure readiness for GED	Groups	Youth/ Adult
Reading Evaluation Adult Diagnosis	READ	Measure reading needs and progress	Indvds.	None Reported
Tests of Adult Basic Educaton	TABE	Measure basic skills achievement	Groups	Child & Adult

ALTERNATIVE (SUPPLEMENTAL) ASSESSMENT METHODS

Methods for assessing learner progress and program outcomes that do not use standardized tests.

Method	Examples of Factors to Assess
Interviews	Reasons for entering program; objectives, pre-program reading, writing, math behavior; post-program changes in behaviors; children's education behaviors pre-post-program
Ratings	Estimates of skill levels by self, teachers, others pre-post program; changes in self-esteem
Portfolio Development	Collections of writing; lists and collections of materials read; lists of real life tasks completed
Performance Samples	Reading aloud; evaluating recordings with peers & teacher; class presentations; community activities